Fair Best Arm Identification with Fixed Confidence

Alessio Russo[*,1], Filippo Vannella[*,2]

2024 Conference on Decision and Control, Milan, Italy

[1]Ericsson AB, [2]Ericsson Research

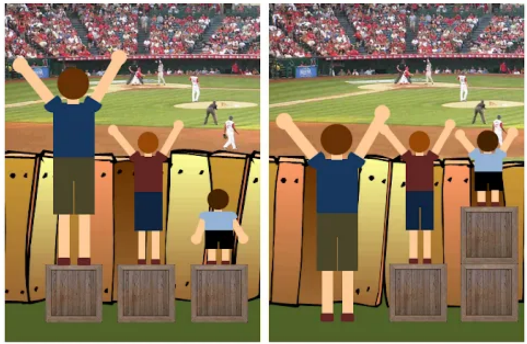
[*] *Equal contribution*. Russo A. is currently at Boston University; Vannella F. is currently at Telefonica.

Introduction and Motivation

Introduction

In recent years, a large body of work has focused on making machine learning systems more **fair** [3].

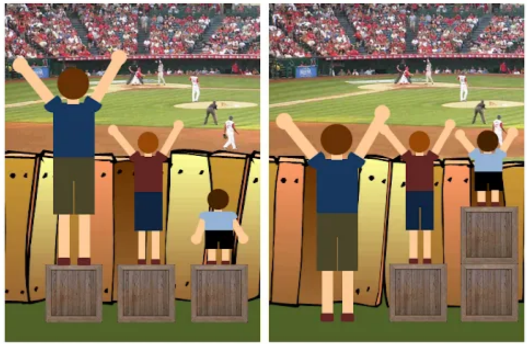
- broader societal shift towards more ethical algorithms.
- Applications in *online advertisement* [21], *recommender systems* [1, 5], *wireless network optimization*.
- Example: in wireless scheduling with multiple QoS classes, fairness ensures users in each class meet their specific performance requirements.



Introduction

In recent years, a large body of work has focused on making machine learning systems more **fair** [3].

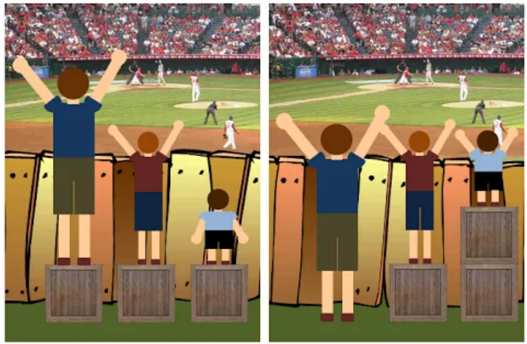
- broader societal shift towards more ethical algorithms.
- Applications in *online advertisement* [21], *recommender systems* [1, 5], *wireless network optimization*.
- **Example**: in wireless scheduling with multiple QoS classes, fairness ensures users in each class meet their specific performance requirements.



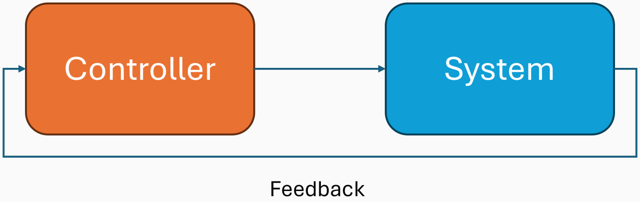
Introduction

In recent years, a large body of work has focused on making machine learning systems more **fair** [3].

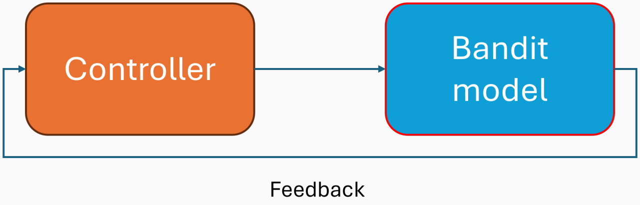
- broader societal shift towards more ethical algorithms.
- Applications in *online advertisement* [21], *recommender systems* [1, 5], *wireless network optimization*.
- **Example**: in wireless scheduling with multiple QoS classes, fairness ensures users in each class meet their specific performance requirements.



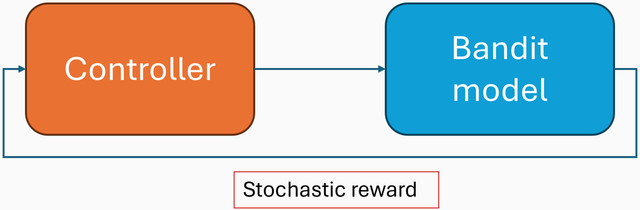
Introduction



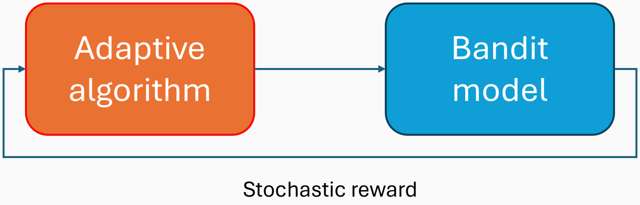
Introduction



Introduction



Introduction



Problem Setting and Related Work

Multi-Armed Bandit Problem

K arms with Gaussian reward distributions $\mathcal{N}(\theta_a, 1)$ with $a \in \{1, \ldots, K\}$.

θ_1

θ_2

θ_3

θ_4

θ_5

- **Sequential**:In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \mathcal{N}(\theta_a, 1)$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \theta_a$ with confidence $\delta \in (0, 1) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.
- **Caveat**: we want to be fair! How? In what way?

Multi-Armed Bandit Problem

K arms with Gaussian reward distributions $\mathcal{N}(\theta_a, 1)$ with $a \in \{1, \ldots, K\}$.

θ_1

θ_2

θ_3

θ_4

θ_5

- **Sequential**:In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \mathcal{N}(\theta_a, 1)$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \theta_a$ with confidence $\delta \in (0,1) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.
- **Caveat**: we want to be fair! How? In what way?

K arms with Gaussian reward distributions $\mathcal{N}(\theta_a, 1)$ with $a \in \{1, \ldots, K\}$.

θ_1

θ_2

θ_3

θ_4

θ_5

- **Sequential**:In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \mathcal{N}(\theta_a, 1)$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \theta_a$ with confidence $\delta \in (0,1) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.
- **Caveat**: we want to be fair! How? In what way?

Multi-Armed Bandit Problem

K arms with Gaussian reward distributions $\mathcal{N}(\theta_a, 1)$ with $a \in \{1, \ldots, K\}$.





θ_1 θ_2 θ_3 θ_4 θ_5

- **Sequential**:In round t the learner pulls arm $a_t \in [K]$ and receives the reward $r_t \sim \mathcal{N}(\theta_a, 1)$.
- **Best Arm Identification objective**: quickly find the optimal arm $a^\star = \arg\max_a \theta_a$ with confidence $\delta \in (0, 1) \Rightarrow$ minimize sample complexity $\mathbb{E}[\tau]$ subject to $\mathbb{P}(\hat{a}_\tau \neq a^\star) \leq \delta$.
  - τ is a random stopping time and $\hat{a}_\tau$ is the estimated best arm at τ.
- **Caveat**: we want to be fair! How? In what way?

Fairness in Bandit Problems

In MAB problems, fairness has been investigated in different bandit settings[12, 14, 11, 9]:

- Many different notions of fairness, most of which fall into the following categories: (I) **pre-specified fairness**; (II) **individual fairness**; (III) counterfactual fairness and (IV) group fairness [7].
- Other important works consider the α-**fairness criterion** [2] for fair resource allocation, which encompasses different fairness criteria when varying the value of the parameter α:
  - **Max-min** fairness: allocates resources as equally as possible.
  - **Proportional fairness**: allocates resources in a proportional manner [19, 20, 18].

Fairness in Bandit Problems

In MAB problems, fairness has been investigated in different bandit settings[12, 14, 11, 9]:

- Many different notions of fairness, most of which fall into the following categories: (I) **pre-specified fairness**; (II) **individual fairness**; (III) counterfactual fairness and (IV) group fairness [7].
- Other important works consider the α-**fairness criterion** [2] for fair resource allocation, which encompasses different fairness criteria when varying the value of the parameter α:
  - **Max-min** fairness: allocates resources as equally as possible.
  - **Proportional fairness**: allocates resources in a proportional manner [19, 20, 18].

Fairness in Bandit Problems

In MAB problems, fairness has been investigated in different bandit settings[12, 14, 11, 9]:

- Many different notions of fairness, most of which fall into the following categories: (I) **pre-specified fairness**; (II) **individual fairness**; (III) counterfactual fairness and (IV) group fairness [7].
- Other important works consider the α-**fairness criterion** [2] for fair resource allocation, which encompasses different fairness criteria when varying the value of the parameter α:
  - **Max-min** fairness: allocates resources as equally as possible.
  - **Proportional fairness**: allocates resources in a proportional manner [19, 20, 18].

Fairness in Bandit Problems

In MAB problems, fairness has been investigated in different bandit settings[12, 14, 11, 9]:

- Many different notions of fairness, most of which fall into the following categories: (I) **pre-specified fairness**; (II) **individual fairness**; (III) counterfactual fairness and (IV) group fairness [7].
- Other important works consider the α-**fairness criterion** [2] for fair resource allocation, which encompasses different fairness criteria when varying the value of the parameter α:
  - **Max-min** fairness: allocates resources as equally as possible.
  - **Proportional fairness**: allocates resources in a proportional manner [19, 20, 18].

Fairness in Bandit Problems

In MAB problems, fairness has been investigated in different bandit settings[12, 14, 11, 9]:

- Many different notions of fairness, most of which fall into the following categories: (I) **pre-specified fairness**; (II) **individual fairness**; (III) counterfactual fairness and (IV) group fairness [7].
- Other important works consider the α-**fairness criterion** [2] for fair resource allocation, which encompasses different fairness criteria when varying the value of the parameter α:
  - **Max-min** fairness: allocates resources as equally as possible.
  - **Proportional fairness**: allocates resources in a proportional manner [19, 20, 18].

Fairness in Bandit Problems

- Selection with pre-specified values of fairness: the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
  - Example of asymptotic fairness [14]:

$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{N_a(T)}{T}\right] \geq p_a \quad \forall a \in [K].$$

    where $N_a(T)$ is the number of times the algorithm selected action a up to time T.
  - [16] define an algorithm to be η-fair if

$$\lfloor p_a t \rfloor - N_a(t) \leq \eta, \forall t \in [T], \forall a \in [K].$$

  - Notably, a fair UCB algorithm guarantees $\mathrm{Reg}(T) \leq (1 + \pi^2/3) \sum_{a \neq a^\star} \Delta_a$ (constant regret).

Fairness in Bandit Problems

- Selection with pre-specified values of fairness: the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
  - Example of asymptotic fairness [14]:

$$\liminf_{T \to \infty} \mathbb{E}\left[\frac{N_a(T)}{T}\right] \geq p_a \quad \forall a \in [K].$$

    where $N_a(T)$ is the number of times the algorithm selected action a up to time T.
  - [16] define an algorithm to be η-fair if

$$\lfloor p_a t \rfloor - N_a(t) \leq \eta, \forall t \in [T], \forall a \in [K].$$

  - Notably, a fair UCB algorithm guarantees $\mathrm{Reg}(T) \leq (1 + \pi^2/3) \sum_{a \neq a^\star} \Delta_a$ (constant regret).

Fairness in Bandit Problems

- Selection with pre-specified values of fairness: the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
  - Example of asymptotic fairness [14]:

$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{N_a(T)}{T}\right] \geq p_a \quad \forall a \in [K].$$

    where $N_a(T)$ is the number of times the algorithm selected action a up to time T.
  - [16] define an algorithm to be η-fair if

$$\lfloor p_a t \rfloor - N_a(t) \leq \eta, \forall t \in [T], \forall a \in [K].$$

  - Notably, a fair UCB algorithm guarantees $\mathrm{Reg}(T) \leq (1 + \pi^2/3) \sum_{a \neq a^\star} \Delta_a$ (constant regret).

Fairness in Bandit Problems

- Selection with pre-specified values of fairness: the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
  - Example of asymptotic fairness [14]:

$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{N_a(T)}{T}\right] \geq p_a \quad \forall a \in [K].$$

    where $N_a(T)$ is the number of times the algorithm selected action a up to time T.
  - [16] define an algorithm to be η-fair if

$$\lfloor p_a t \rfloor - N_a(t) \leq \eta, \forall t \in [T], \forall a \in [K].$$

  - Notably, a fair UCB algorithm guarantees $\mathrm{Reg}(T) \leq (1 + \pi^2/3) \sum_{a \neq a^\star} \Delta_a$ (constant regret).

Fairness in Bandit Problems

- Selection with pre-specified values of fairness : the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
- Individual and proportional fairness: requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [12] – these are model-dependent constraints.

Fairness in Bandit Problems

- Selection with pre-specified values of fairness : the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
- Individual and proportional fairness: requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [12] – these are model-dependent constraints.
- α-fairness criterion: find a policy maximizing the α-criterion

$$f_\alpha(\theta) = \begin{cases} \frac{\theta^{1-\alpha}}{1-\alpha} & \alpha \in [0,1) \cup (1,\infty), \\ \log(\theta) & \alpha = 1. \end{cases}$$

  - For $\alpha \to \infty$ we obtain *max-min* fairness: allocate resources equally.
  - For $\alpha = 0$ we obtain the greedy solution.
  - For $\alpha = 1$ we obtain the *proportional fair* solution, which is part of the more general notion of *individual fairness*.

Fairness in Bandit Problems

- Selection with pre-specified values of fairness : the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
- Individual and proportional fairness: requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [12] – these are model-dependent constraints.
- α-fairness criterion: find a policy maximizing the α-criterion

$$f_\alpha(\theta) = \begin{cases} \frac{\theta^{1-\alpha}}{1-\alpha} & \alpha \in [0,1) \cup (1,\infty), \\ \log(\theta) & \alpha = 1. \end{cases}$$

  - For $\alpha \to \infty$ we obtain *max-min* fairness: allocate resources equally.
  - For $\alpha = 0$ we obtain the greedy solution.
  - For $\alpha = 1$ we obtain the *proportional fair* solution, which is part of the more general notion of *individual fairness*.

Fairness in Bandit Problems

- Selection with pre-specified values of fairness : the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
- Individual and proportional fairness: requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [12] – these are model-dependent constraints.
- α-fairness criterion: find a policy maximizing the α-criterion

$$f_\alpha(\theta) = \begin{cases} \frac{\theta^{1-\alpha}}{1-\alpha} & \alpha \in [0,1) \cup (1,\infty), \\ \log(\theta) & \alpha = 1. \end{cases}$$

  - For $\alpha \to \infty$ we obtain *max-min* fairness: allocate resources equally.
  - For $\alpha = 0$ we obtain the greedy solution.
  - For $\alpha = 1$ we obtain the *proportional fair* solution, which is part of the more general notion of *individual fairness*.

Fairness in Bandit Problems

- Selection with pre-specified values of fairness : the rate at which an algorithm selects an arm a stays within a pre-specified range $[p_0, p_1]$. These constraints are, in general, model-agnostic.
- Individual and proportional fairness: requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [12] – these are model-dependent constraints.
- α-fairness criterion: find a policy maximizing the α-criterion

$$f_\alpha(\theta) = \begin{cases} \frac{\theta^{1-\alpha}}{1-\alpha} & \alpha \in [0,1) \cup (1,\infty), \\ \log(\theta) & \alpha = 1. \end{cases}$$

  - For $\alpha \to \infty$ we obtain *max-min* fairness: allocate resources equally.
  - For $\alpha = 0$ we obtain the greedy solution.
  - For $\alpha = 1$ we obtain the *proportional fair* solution, which is part of the more general notion of *individual fairness*.

Fair Best Arm Identification - Constraints type

What we study

1. Pre-specified constraints: the selection rate at the random stopping time τ, needs be larger than some *pre-specified* value $p_a \in [0,1]$:

$$\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a, \forall a \in [K].$$

2. θ-dependent constraints: asymptotically, as $\delta \to 0$, the selection rate at the stopping time τ needs to be larger than some θ-dependent continuous function $p_a(\theta) : \mathbb{R}^K \to [0,1]$:

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a(\theta), \forall a \in [K].$$

3. Example: $p(\theta) = p_0 \cdot \texttt{softmax}(\theta)$ for some $p_0 \in [0,1]$ (with $p_a(\theta) = (p(\theta))_a$).

Fair Best Arm Identification - Constraints type

What we study

1. Pre-specified constraints: the selection rate at the random stopping time τ, needs be larger than some *pre-specified* value $p_a \in [0, 1]$:

$$\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a, \forall a \in [K].$$

2. θ-dependent constraints: asymptotically, as $\delta \to 0$, the selection rate at the stopping time τ needs to be larger than some θ-dependent continuous function $p_a(\theta) : \mathbb{R}^K \to [0, 1]$:

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a(\theta), \forall a \in [K].$$

3. Example: $p(\theta) = p_0 \cdot \texttt{softmax}(\theta)$ for some $p_0 \in [0, 1]$ (with $p_a(\theta) = (p(\theta))_a$).

Fair Best Arm Identification - Constraints type

What we study

1. Pre-specified constraints: the selection rate at the random stopping time τ, needs be larger than some *pre-specified* value $p_a \in [0, 1]$:

$$\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a, \forall a \in [K].$$

2. θ-dependent constraints: asymptotically, as $\delta \to 0$, the selection rate at the stopping time τ needs to be larger than some θ-dependent continuous function $p_a(\theta) : \mathbb{R}^K \to [0, 1]$:

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a(\theta), \forall a \in [K].$$

3. Example: $p(\theta) = p_0 \cdot \texttt{softmax}(\theta)$ for some $p_0 \in [0, 1]$ (with $p_a(\theta) = (p(\theta))_a$).

p-fair δ-Probably Correct Algorithm

Definition

- An algorithm is p-fair δ-PC (Probably Correct) if for all $\theta \in \Theta$, $\delta \in (0, 1/2)$, it satisfies

$$(i)\ \frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a, \forall a \in [K], \quad (ii)\ \mathbb{P}_\theta(\hat{a}_\tau \neq a^\star) \leq \delta, \quad (iii)\ \mathbb{P}_\theta(\tau < \infty) = 1. \tag{1}$$

- Similarly, we say an algorithm is asymptotically $p(\theta)$-fair δ-PC if it satisfies

$$(i)\ \liminf_{\delta \to 0} \frac{\mathbb{E}_\theta[N_a(\tau_\delta)]}{\mathbb{E}_\theta[\tau_\delta]} \geq p_a(\theta), \forall a \in [K], \quad (ii)\ \mathbb{P}_\theta(\hat{a}_\tau \neq a^\star) \leq \delta, \quad (iii)\ \mathbb{P}_\theta(\tau < \infty) = 1. \tag{2}$$

p-fair δ-Probably Correct Algorithm

Definition

- An algorithm is p-fair δ-PC (Probably Correct) if for all $\theta \in \Theta$, $\delta \in (0, 1/2)$, it satisfies

$$(i)\ \frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]} \geq p_a, \forall a \in [K], \quad (ii)\ \mathbb{P}_\theta(\hat{a}_\tau \neq a^\star) \leq \delta, \quad (iii)\ \mathbb{P}_\theta(\tau < \infty) = 1. \tag{1}$$

- Similarly, we say an algorithm is asymptotically $p(\theta)$-fair δ-PC if it satisfies

$$(i)\ \liminf_{\delta \to 0} \frac{\mathbb{E}_\theta[N_a(\tau_\delta)]}{\mathbb{E}_\theta[\tau_\delta]} \geq p_a(\theta), \forall a \in [K], \quad (ii)\ \mathbb{P}_\theta(\hat{a}_\tau \neq a^\star) \leq \delta, \quad (iii)\ \mathbb{P}_\theta(\tau < \infty) = 1. \tag{2}$$

Main Result: Sample Complexity Lower Bound

Sample Complexity Lower Bound

- **Define the characteristic time**

$$T_p^\star = 2 \inf_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2},$$

where $\Sigma_p = \{w \geq p : \sum_{a \in [K]} w_a = 1\}$ is the clipped simplex.

  - w_a is the optimal static rate at which the agent should select arm a.
  - $\Delta_a = \theta_{a^\star} - \theta$ is the sub-optimality gap in a.
- The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm.
  - It is derived using hypothesis-testing argument. Consider a trajectory $\tau = (a_1, r_1, \ldots, a_t, r_t)$: is this data generated using the true model θ or a *confusing* one θ'?
  - Construct the log-likelihood ratio $L_t = \log \frac{\mathrm{d}\mathbb{P}_\theta(\tau)}{\mathrm{d}\mathbb{P}_{\theta'}(\tau)}$.
  - Find θ' by minimizing

$$\min_{\theta'} \mathbb{E}_\theta[L_\tau] = \min_{\theta'} \sum_a \mathbb{E}_\theta[N_a(\tau)] \mathrm{KL}(P_{\theta_a}, P_{\theta'_a}) = \mathbb{E}_\theta[\tau] \min_{\theta'} \sum_a \underbrace{\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]}}_{=: w_a} \mathrm{KL}(P_{\theta_a}, P_{\theta'_a})$$

over models θ' that admits a different optimal action.

Sample Complexity Lower Bound

- **Define the characteristic time**

$$T_p^\star = 2 \inf_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2},$$

  where $\Sigma_p = \{w \geq p : \sum_{a \in [K]} w_a = 1\}$ is the clipped simplex.
  - w_a is the optimal static rate at which the agent should select arm a.
  - $\Delta_a = \theta_{a^\star} - \theta$ is the sub-optimality gap in a.
- The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm.
  - It is derived using hypothesis-testing argument. Consider a trajectory $\tau = (a_1, r_1, \dots, a_t, r_t)$: is this data generated using the true model θ or a *confusing* one θ'?
  - Construct the log-likelihood ratio $L_t = \log \frac{\mathrm{d}\mathbb{P}_\theta(\tau)}{\mathrm{d}\mathbb{P}_{\theta'}(\tau)}$.
  - Find θ' by minimizing

$$\min_{\theta'} \mathbb{E}_\theta[L_\tau] = \min_{\theta'} \sum_a \mathbb{E}_\theta[N_a(\tau)] \mathrm{KL}(P_{\theta_a}, P_{\theta'_a}) = \mathbb{E}_\theta[\tau] \min_{\theta'} \sum_a \underbrace{\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]}}_{=: w_a} \mathrm{KL}(P_{\theta_a}, P_{\theta'_a})$$

    over models θ' that admits a different optimal action.

Sample Complexity Lower Bound

- **Define the characteristic time**

$$T_p^\star = 2 \inf_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2},$$

  where $\Sigma_p = \{w \geq p : \sum_{a \in [K]} w_a = 1\}$ is the clipped simplex.
  - w_a is the optimal static rate at which the agent should select arm a.
  - $\Delta_a = \theta_{a^\star} - \theta$ is the sub-optimality gap in a.
- The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm.
  - It is derived using hypothesis-testing argument. Consider a trajectory $\tau = (a_1, r_1, \ldots, a_t, r_t)$: is this data generated using the true model θ or a *confusing* one θ'?
  - Construct the log-likelihood ratio $L_t = \log \frac{\mathrm{d}\mathbb{P}_\theta(\tau)}{\mathrm{d}\mathbb{P}_{\theta'}(\tau)}$.
  - Find θ' by minimizing

$$\min_{\theta'} \mathbb{E}_\theta[L_\tau] = \min_{\theta'} \sum_a \mathbb{E}_\theta[N_a(\tau)] \mathrm{KL}(P_{\theta_a}, P_{\theta'_a}) = \mathbb{E}_\theta[\tau] \min_{\theta'} \sum_a \underbrace{\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]}}_{=: w_a} \mathrm{KL}(P_{\theta_a}, P_{\theta'_a})$$

  over models θ' that admits a different optimal action.

Sample Complexity Lower Bound

- **Define the characteristic time**

$$T_p^\star = 2 \inf_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2},$$

  where $\Sigma_p = \{w \geq p : \sum_{a \in [K]} w_a = 1\}$ is the clipped simplex.
  - w_a is the optimal static rate at which the agent should select arm a.
  - $\Delta_a = \theta_{a^\star} - \theta$ is the sub-optimality gap in a.
- The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm.
  - It is derived using hypothesis-testing argument. Consider a trajectory $\tau = (a_1, r_1, \ldots, a_t, r_t)$: is this data generated using the true model θ or a *confusing* one θ'?
  - Construct the log-likelihood ratio $L_t = \log \frac{\mathrm{d}\mathbb{P}_\theta(\tau)}{\mathrm{d}\mathbb{P}_{\theta'}(\tau)}$.
  - Find θ' by minimizing

$$\min_{\theta'} \mathbb{E}_\theta[L_\tau] = \min_{\theta'} \sum_a \mathbb{E}_\theta[N_a(\tau)] \mathrm{KL}(P_{\theta_a}, P_{\theta'_a}) = \mathbb{E}_\theta[\tau] \min_{\theta'} \sum_a \underbrace{\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]}}_{=: w_a} \mathrm{KL}(P_{\theta_a}, P_{\theta'_a})$$

    over models θ' that admits a different optimal action.

Sample Complexity Lower Bound

- **Define the characteristic time**

$$T_p^\star = 2 \inf_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2},$$

where $\Sigma_p = \{w \geq p : \sum_{a \in [K]} w_a = 1\}$ is the clipped simplex.

  - w_a is the optimal static rate at which the agent should select arm a.
  - $\Delta_a = \theta_{a^\star} - \theta$ is the sub-optimality gap in a.
- The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm.
  - It is derived using hypothesis-testing argument. Consider a trajectory $\tau = (a_1, r_1, \dots, a_t, r_t)$: is this data generated using the true model θ or a *confusing* one θ'?
  - Construct the log-likelihood ratio $L_t = \log \frac{\mathrm{d}\mathbb{P}_\theta(\tau)}{\mathrm{d}\mathbb{P}_{\theta'}(\tau)}$.
  - Find θ' by minimizing

$$\min_{\theta'} \mathbb{E}_\theta[L_\tau] = \min_{\theta'} \sum_a \mathbb{E}_\theta[N_a(\tau)] \mathrm{KL}(P_{\theta_a}, P_{\theta'_a}) = \mathbb{E}_\theta[\tau] \min_{\theta'} \sum_a \underbrace{\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]}}_{=: w_a} \mathrm{KL}(P_{\theta_a}, P_{\theta'_a})$$

over models θ' that admits a different optimal action.

- **Define the characteristic time**

$$T_p^\star = 2 \inf_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2},$$

  where $\Sigma_p = \{w \geq p : \sum_{a \in [K]} w_a = 1\}$ is the clipped simplex.
  - w_a is the optimal static rate at which the agent should select arm a.
  - $\Delta_a = \theta_{a^\star} - \theta$ is the sub-optimality gap in a.
- The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm.
  - It is derived using hypothesis-testing argument. Consider a trajectory $\tau = (a_1, r_1, \ldots, a_t, r_t)$: is this data generated using the true model θ or a *confusing* one θ'?
  - Construct the log-likelihood ratio $L_t = \log \frac{\mathrm{d}\mathbb{P}_\theta(\tau)}{\mathrm{d}\mathbb{P}_{\theta'}(\tau)}$.
  - Find θ' by minimizing

$$\min_{\theta'} \mathbb{E}_\theta[L_\tau] = \min_{\theta'} \sum_a \mathbb{E}_\theta[N_a(\tau)] \mathrm{KL}(P_{\theta_a}, P_{\theta'_a}) = \mathbb{E}_\theta[\tau] \min_{\theta'} \sum_a \underbrace{\frac{\mathbb{E}_\theta[N_a(\tau)]}{\mathbb{E}_\theta[\tau]}}_{=: w_a} \mathrm{KL}(P_{\theta_a}, P_{\theta'_a})$$

    over models θ' that admits a different optimal action.

Sample Complexity Lower Bound

- **Define the characteristic time**

$$T_p^\star = 2 \inf_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2},$$

where $\Sigma_p = \{w \geq p : \sum_{a \in [K]} w_a = 1\}$ is the clipped simplex.

- The characteristic time $T_p^\star$ represents the difficulty of identifying the best arm.

Theorem

- *Any p-fair δ-PAC algorithm satisfies*

$$\frac{\mathbb{E}_\theta[\tau]}{\log(1/2.4\delta)} \geq T_p^\star \quad \forall \theta \in \Theta.$$

- *Any asymptotically $p(\theta)$-fair δ-PAC algorithm satisfies*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\theta[\tau]}{\log(1/\delta)} \geq T_p^\star \quad \forall \theta \in \Theta.$$

Cost of Fairness

Lemma

For a set of fairness constraints $p = (p_a)_{a \in [K]}$, *and for all* $\theta \in \Theta$, *we have that*

$$1 \leq \frac{T_p^\star}{T_0^\star} \leq O\left(\min\left(\frac{1}{1 - p_{\mathrm{sum}}}, \frac{1}{K p_{\min}}\right)\right). \tag{3}$$

where $p_{\mathrm{sum}} = \sum_a p_a$ *and* $p_{\min} = \min_{a : p_a > 0} p_a$.

- T_0 denotes the characteristic time with $p = 0$ (no fairness constraints).
- Price of fairness typically scales as $(1 - p_{\mathrm{sum}})^{-1}$ or $(p_{\min})^{-1}$!

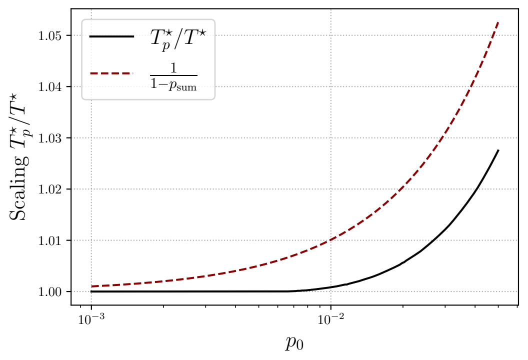
Cost of Fairness: Example

$$1 \leq \frac{T_p^\star}{T_0^\star} \leq O\left(\min\left(\frac{1}{1 - p_{\text{sum}}}, \frac{1}{K p_{\min}}\right)\right). \tag{4}$$

where $p_{\text{sum}} = \sum_a p_a$ and $p_{\min} = \min_{a: p_a > 0} p_a$.

Antagonistic scenario:

- $p_a(\theta) = K p_0 \frac{\Delta_a}{\sum_b \Delta_b}$, for $p_0 \in [0, 1/K]$.
- $p_a(\theta) \propto \Delta_a \Rightarrow$ larger for sub-optimal arms.
- Small $p_0 \Rightarrow p_{\min}^{-1} > (1 - p_{\text{sum}})^{-1}$, with $(p_{\min})^{-1} = O(1/p_0)$.
- $T_p^\star / T_0^\star$ scales according to $(1 - p_{\text{sum}})^{-1}$.



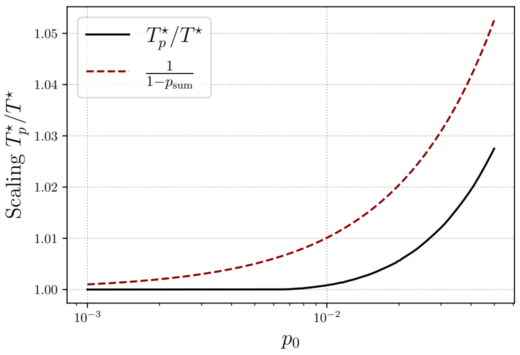
Cost of Fairness: Example

$$1 \leq \frac{T_p^\star}{T_0^\star} \leq O\left(\min\left(\frac{1}{1 - p_{\text{sum}}}, \frac{1}{K p_{\min}}\right)\right). \quad (4)$$

where $p_{\text{sum}} = \sum_a p_a$ and $p_{\min} = \min_{a: p_a > 0} p_a$.

Antagonistic scenario:

- $p_a(\theta) = K p_0 \frac{\Delta_a}{\sum_b \Delta_b}$, for $p_0 \in [0, 1/K]$.
- $p_a(\theta) \propto \Delta_a \Rightarrow$ larger for sub-optimal arms.
- Small $p_0 \Rightarrow p_{\min}^{-1} > (1 - p_{\text{sum}})^{-1}$, with $(p_{\min})^{-1} = O(1/p_0)$.
- $T_p^\star / T_0^\star$ scales according to $(1 - p_{\text{sum}})^{-1}$.



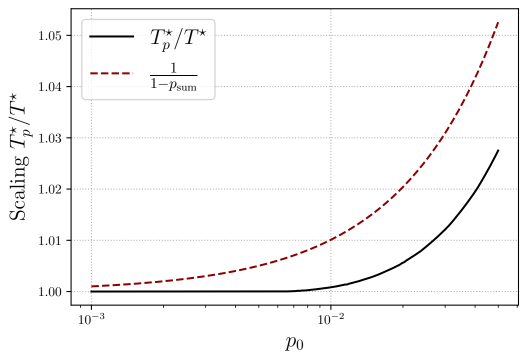
Cost of Fairness: Example

$$1 \leq \frac{T_p^\star}{T_0^\star} \leq O\left(\min\left(\frac{1}{1-p_{\text{sum}}}, \frac{1}{Kp_{\min}}\right)\right). \quad (4)$$

where $p_{\text{sum}} = \sum_a p_a$ and $p_{\min} = \min_{a:p_a>0} p_a$.

Antagonistic scenario:

- $p_a(\theta) = Kp_0 \frac{\Delta_a}{\sum_b \Delta_b}$, for $p_0 \in [0, 1/K]$.
- $p_a(\theta) \propto \Delta_a \Rightarrow$ larger for sub-optimal arms.
- Small $p_0 \Rightarrow p_{\min}^{-1} > (1-p_{\text{sum}})^{-1}$, with $(p_{\min})^{-1} = O(1/p_0)$.
- $T_p^\star / T_0^\star$ scales according to $(1-p_{\text{sum}})^{-1}$.



Method: Fair Track and Stop (F-TaS)

F-TaS: Fair Track and Stop

We propose F-TaS, an (asymptotically) p-fair and δ-PAC algorithm. It consists of (i) a sampling rule and (ii) a stopping rule.

F-TaS: Sampling Rule

Sampling Rule

The fundamental idea is that the lower bound provides you w, the optimal way to sample actions (i.e., sample $a \sim w$), where w is computed according to

$$T_p^\star = 2 \min_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2}$$

- But we don't know Δ_a! We plug-in the estimate $\Delta_a(t)$ at time t:

$$w_p^\star(t) = \arg\min_{w \in \Sigma_p} \max_{a \neq a_t^\star} \frac{w_a^{-1} + w_{a_t^\star}(t)^{-1}}{\Delta_a(t)^2}$$

- Σ_p depends on $\theta(t)$ (the estimate at time t of the means) in the θ-dependent constraints.
- To ensure $\theta(t) \to \theta$ we mix $w_p^\star(t)$ with a constant policy $\pi_c = (\pi_{c,a})_{a\in[K]}$, using a parameter ϵ_t (*forced exploration policy*).

F-TaS: Sampling Rule

Sampling Rule

The fundamental idea is that the lower bound provides you w, the optimal way to sample actions (i.e., sample $a \sim w$), where w is computed according to

$$T_p^\star = 2 \min_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2}$$

- But we don't know Δ_a! We plug-in the estimate $\Delta_a(t)$ at time t:

$$w_p^\star(t) = \arg\min_{w \in \Sigma_p} \max_{a \neq a_t^\star} \frac{w_a^{-1} + w_{a_t^\star}(t)^{-1}}{\Delta_a(t)^2}$$

- Σ_p depends on $\theta(t)$ (the estimate at time t of the means) in the θ-dependent constraints.
- To ensure $\theta(t) \to \theta$ we mix $w_p^\star(t)$ with a constant policy $\pi_c = (\pi_{c,a})_{a \in [K]}$, using a parameter ϵ_t (*forced exploration policy*).

F-TaS: Sampling Rule

Sampling Rule

The fundamental idea is that the lower bound provides you w, the optimal way to sample actions (i.e., sample $a \sim w$), where w is computed according to

$$T_p^\star = 2 \min_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2}$$

- But we don't know Δ_a! We plug-in the estimate $\Delta_a(t)$ at time t:

$$w_p^\star(t) = \arg\min_{w \in \Sigma_p} \max_{a \neq a_t^\star} \frac{w_a^{-1} + w_{a_t^\star}(t)^{-1}}{\Delta_a(t)^2}$$

- Σ_p depends on $\theta(t)$ (the estimate at time t of the means) in the θ-dependent constraints.
- To ensure $\theta(t) \to \theta$ we mix $w_p^\star(t)$ with a constant policy $\pi_c = (\pi_{c,a})_{a \in [K]}$, using a parameter ϵ_t (*forced exploration policy*).

F-TaS: Sampling Rule

Sampling Rule

The fundamental idea is that the lower bound provides you w, the optimal way to sample actions (i.e., sample $a \sim w$), where w is computed according to

$$T_p^\star = 2 \min_{w \in \Sigma_p} \max_{a \neq a^\star} \frac{w_a^{-1} + w_{a^\star}^{-1}}{\Delta_a^2}$$

- But we don't know Δ_a! We plug-in the estimate $\Delta_a(t)$ at time t:

$$w_p^\star(t) = \arg\min_{w \in \Sigma_p} \max_{a \neq a_t^\star} \frac{w_a^{-1} + w_{a_t^\star}(t)^{-1}}{\Delta_a(t)^2}$$

- Σ_p depends on $\theta(t)$ (the estimate at time t of the means) in the θ-dependent constraints.
- To ensure $\theta(t) \to \theta$ we mix $w_p^\star(t)$ with a constant policy $\pi_c = (\pi_{c,a})_{a \in [K]}$, using a parameter ϵ_t (*forced exploration policy*).

F-TaS: Stopping Rule

Stopping Rule

The stopping rule should stop as soon as we are confident of the best arm. Stop as soon as

$$t \,\tilde{>}\, T_p^\star(t) \log\left(\frac{1+\log(t)}{\delta}\right),$$

where $T_p^\star(t)$ is the estimate at time t of $T_p^\star$, computed as

$$T_p^\star(t) := 2 \max_{a \neq a_t^\star} \frac{w_a(t)^{-1} + w_{a_t^\star}(t)^{-1}}{\Delta_a(t)^2}, \; w_a(t) := \frac{N_a(t)}{t}.$$

F-TaS: Fair Track and Stop

Algorithm 1 F-TaS

Input: Fairness vector $p = (p_a)_{a\in[K]}$; confidence δ; forced exploration schedule $(\epsilon_t)_t$.
Set $t \leftarrow 1$
while $t \le T_p^\star(t)\log\left(\frac{1+\log(t)}{\delta}\right)$ **do**
  Compute $w_p^\star(t) = \min_{w\in\Sigma_p}\max_{a\neq a_t^\star}\frac{w_a(t)^{-1}+w_{a_t^\star}(t)^{-1}}{\Delta_a(t)^2}$ and set $\pi(t) \leftarrow (1-\epsilon_t)w_p^\star(t)+\epsilon_t\pi_c$
  Select $a_t \sim \pi(t)$ and observe reward r_t
  Update statistics $\hat\theta(t), N_a(t)$ and set $t \leftarrow t+1$
end while
Return $\hat a_\tau = \arg\max_a \hat\theta_a(\tau)$

F-TaS: Guarantees

Theorem

- F-TaS *is p-fair (resp. asymptotically $p(\theta)$-fair) and δ-PAC.*
- *For all $\delta \in (0, 1/2)$,* F-TaS *has a finite expected sample complexity $\mathbb{E}_\theta[\tau_\delta] < \infty$, and it satisfies:*

  (1) *Almost sure asymptotic optimality:*

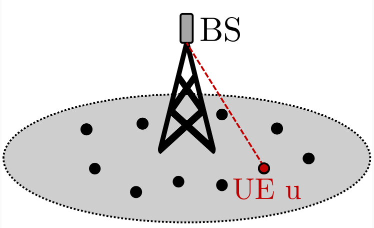
$$\mathbb{P}_\theta\left(\limsup_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \le T_p^\star\right) = 1,$$

  (2) *Asymptotic optimality in expectation:*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\theta[\tau]}{\log(1/\delta)} \le T_p^\star.$$

Numerical Results

Wireless Scheduling: Model



Base Station (BS) and a set of K User Equipments (UEs).

Model:

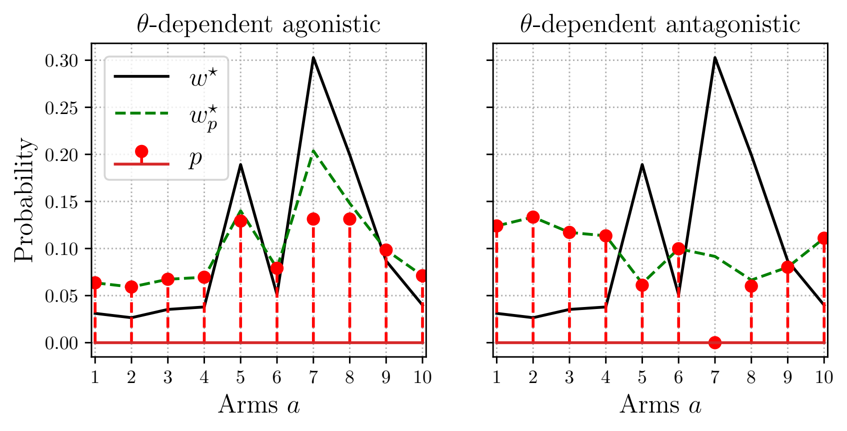
- At each round, $t \geq 1$, the BS selects a single UE out of the K to be scheduled for transmission.
- the BS represents the learner, and the set of UEs $[K]$ represents the various arms.
- The reward at round t is defined as the sum throughput across UEs in the cell, i.e., $r_t = \sum_{u \in [K]} T_{u,t} \mathbf{1}_{\{a_t = u\}}$.
- We compare F-TaS with Track and Stop (TaS [8]) and Uniform Fair, an algorithm selecting an arm a in round t with probability $p_a(\hat{\theta}(t)) + (1 - p_{\text{sum}}(\hat{\theta}(t)))/K$.

Wireless Scheduling: Fairness

Settings: we focus on two settings to analyze how p impacts exploration:

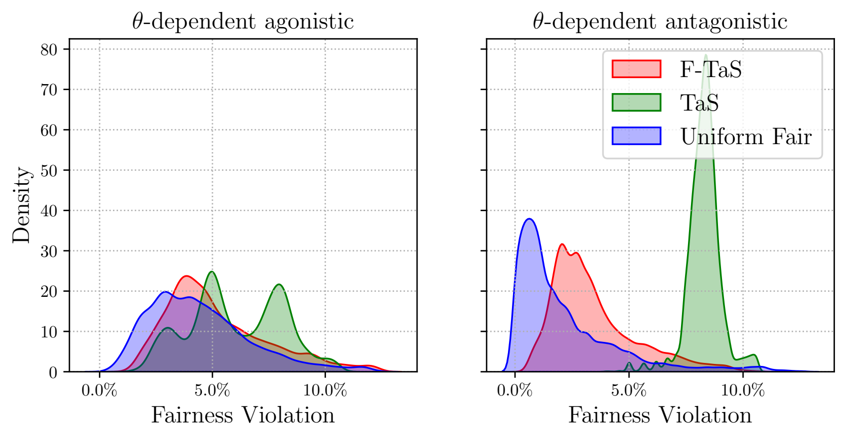
1. *agonistic fairness*: promotes exploration
2. *antagonistic fairness*: inhibits exploration.

Wireless Scheduling: Optimal Allocations



Optimal action selection frequencies: $w^\star$ is the optimal solution to $T_0^\star$ (i.e., with $p = 0$).

Wireless Scheduling: Fairness Violation, θ-dependent scenario



Fairness violation: $\rho(t) = \max(0, \max_a p_a(\theta) - N_a(t)/t)$.

Conclusions

Conclusions

Lot of interest in making learning algorithms more fair:

- Fairness can help with exploration, or regret minimization (e.g., constant regret).
- How do we achieve fairness in more complex adaptive systems?
- How to extend to general Markov Decision Processes?
- Find the code here
  `https://github.com/rssalessio/fair-best-arm-identification`

Thank you for listening!

References i

Kaito Ariu, Narae Ryu, Se-Young Yun, and Alexandre Proutière.
Regret in online recommendation systems.
In *Proc. of NeurIPS*, 2020.

Anthony B Atkinson et al.
On the measurement of inequality.
Journal of economic theory, 1970.

Simon Caton and Christian Haas.
Fairness in machine learning: A survey.
ACM Computing Surveys, 2020.

L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi.
Controlling polarization in personalization: An algorithmic framework.
In *Proc. of the conference on fairness, accountability, and transparency*, 2019.

References ii

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire.
Contextual bandits with linear payoff functions.
In *Proc. of AISTATS*, 2011.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel.
Fairness through awareness.
In *Proc. of the 3rd innovations in theoretical computer science conference*, 2012.

Pratik Gajane, Akrati Saxena, Maryam Tavakol, George Fletcher, and Mykola Pechenizkiy.
Survey on fair reinforcement learning: Theory and practice.
arXiv preprint arXiv:2205.10032, 2022.

Aurélien Garivier and Emilie Kaufmann.
Optimal best arm identification with fixed confidence.
In *Proc. of COLT*. PMLR, 2016.

References iii

Riccardo Grazzi, Arya Akhavan, John IF Falk, Leonardo Cella, and Massimiliano Pontil.
Group meritocratic fairness in linear contextual bandits.
In *Proc. of NeurIPS*, 2022.

Wen Huang, Kevin Labille, Xintao Wu, Dongwon Lee, and Neil Heffernan.
Achieving user-side fairness in contextual bandits.
Human-Centric Intelligent Systems, 2022.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth.
Fairness in reinforcement learning.
In *ICML*, 2017.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth.
Fairness in learning: Classic and contextual bandits.
In *Proc. of NeurIPS*, 2016.

- Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva.
  Counterfactual fairness.
  In *Proc. of NeurIPS*, 2017.
- Fengjiao Li, Jia Liu, and Bo Ji.
  Combinatorial sleeping bandits with fairness constraints.
  IEEE Transactions on Network Science and Engineering, 2019.
- Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes.
  Calibrated fairness in bandits.
  arXiv preprint arXiv:1707.01875, 2017.
- Vitshakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari.
  Achieving fairness in the stochastic multi-armed bandit problem.
  In *JMLR*, 2021.

References v

Candice Schumann, Zhi Lang, Nicholas Mattei, and John P Dickerson.
Group fairness in bandit arm selection.
arXiv preprint arXiv:1912.03802, 2019.

Mohammad Sadegh Talebi and Alexandre Proutiere.
Learning proportionally fair allocations with low regret.
Proc. of the ACM on Measurement and Analysis of Computing Systems, 2018.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims.
Fairness of exposure in stochastic bandits.
In *ICML*, 2021.

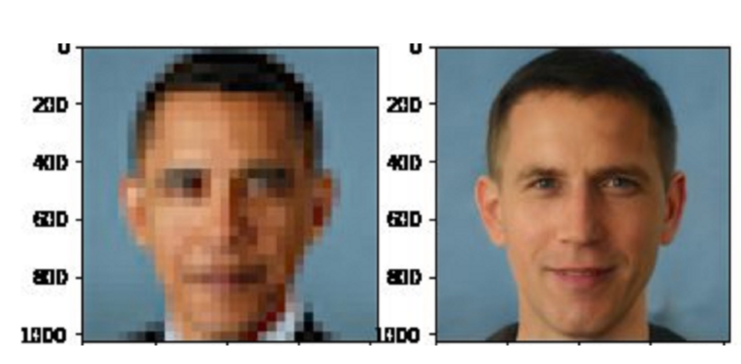
Tianyu Wang and Cynthia Rudin.
Bandit learning for proportionally fair allocations.
`https://wangt1anyu.github.io/papers/prop-fair-bandit.pdf`, 2021.

Min Xu, Tao Qin, and Tie-Yan Liu.
Estimation bias in multi-armed bandit algorithms for search advertising.
In *Proc. of NeurIPS*, 2013.

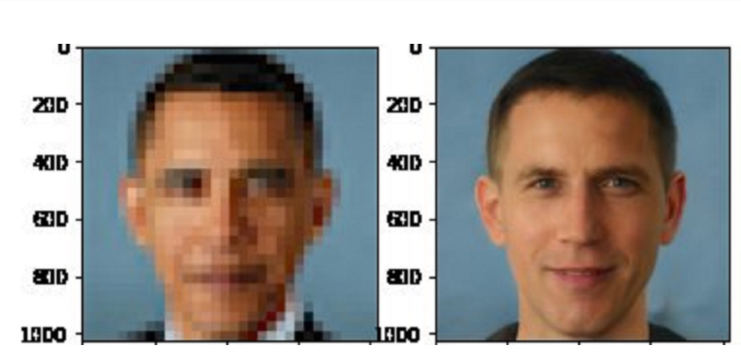
Xueru Zhang and Mingyan Liu.
Fairness in learning-based sequential decision algorithms: A survey.
In *Handbook of Reinforcement Learning and Control*. Springer, 2021.

Appendix

Introduction



Example: ML models may be biased against minorities



Example: ML models may be biased against minorities

How do we make the control loop fair?

Fairness in Bandit Problems

- Traditional Bandit algorithms are not fair.
- Several works investigate regret minimization [12, 14, 16, 22]:

$$\mathrm{Reg}(T) = \theta_{a^\star} T - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

- However, this aspect remains largely unexplored within the problem of Best Arm Identification (BAI)

Fairness in Bandit Problems

- Traditional Bandit algorithms are not fair.
- Several works investigate regret minimization [12, 14, 16, 22]:

$$\mathrm{Reg}(T) = \theta_{a^\star} T - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

- However, this aspect remains largely unexplored within the problem of Best Arm Identification (BAI)

Fairness in Bandit Problems

- Traditional Bandit algorithms are not fair.
- Several works investigate regret minimization [12, 14, 16, 22]:

$$\mathrm{Reg}(T) = \theta_{a^\star} T - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$

- However, this aspect remains largely unexplored within the problem of Best Arm Identification (BAI)

Related Work - Extended

- *Individual fairness* [6, 12] requires a system to make comparable decisions for similar individuals, and the constraints could be based on similarity or merit [20, 15].
- *Selection with pre-specified range* [4, 16, 14] simply demands that the rate, or probability, at which an algorithm selects an arm stays within a pre-specified range.
- *Group fairness* imposes constraints based on some statistical parity across subgroups [7]. For example, in [17] divide arms into several subgroups, and ensure that the probability of pulling an arm is constant given the group membership. In contextual bandit problems, one can ensure fairness among different contexts, as in [10] or between groups similarly to the non-contextual setting [9].
- In [13] the authors study the concept of *counterfactual fairness*. Their definition captures the idea that a decision is fair towards an individual if it is fair also in an alternative situation where the individual belong to a different group while keeping all the other important variables unchanged.

Sample-path fairness

- An alternative definition of fairness could consider constraints of the type $\mathbb{E}_\theta \left[N_a(\tau_\delta)/\tau_\delta \right] \geq p_a(\theta)$.
- We refer to this as "*sample-path fairness*" as it evaluates fairness on each sample path.

Corollary

F-TaS *is sample-path* p*-fair (resp.* $p(\theta)$*-fair), i.e., it satisfies*

- $\mathbb{E}_\theta \left[N_a(\tau_\delta)/\tau_\delta \right] \geq p_a(\theta), \forall a \in [K]$.
- $\liminf_{\delta \to 0} \mathbb{E}_\theta \left[\frac{N_a(\tau_\delta)}{\tau_\delta} \right] \geq p_a(\theta)$.

The idea is to use the fact that

$$\frac{\mathbb{E}_\theta[N_a(\tau_\delta)]}{\mathbb{E}_\theta[\tau_\delta]} \leq \mathbb{E}_\theta \left[N_a(\tau_\delta)/\tau_\delta \right] - \mathrm{Cov}_\theta \left(N_a(\tau_\delta), 1/\tau_\delta \right).$$

and show that the covariance term tends to 0 as $\delta \to 0$.

Forced Exploration Policy

The constant policy π_c, and the value of ϵ_t depend on the type of fairness constraint:

- Pre-specified constraints: Let $K_0 = |\{a \in [K] : p_a = 0\}|$ be the number of arms for which $p_a = 0$. In the simple case that $K_0 = 0$, we set $\pi_{c,a} = p_a + (1 - p_{\text{sum}})/K$. Otherwise we set $\epsilon_t = 1/(2\sqrt{t})$, and define π_c as

$$\pi_{c,a} = \begin{cases} p_a & p_a > 0 \\ \frac{1-p_{\text{sum}}}{K_0} & \text{otherwise.} \end{cases}$$

  - this constraint induces a linear exploration rate and hence we do not require any additional forced exploration.
- θ-dependent constraints: in this case, we select $\pi_{c,a} = 1/K$, i.e., a uniform policy for all $a \in [K]$, and we set $\epsilon_t = 1/(2\sqrt{t})$.

Wireless Scheduling: Fairness

Settings: we focus on two settings to analyze how p impacts exploration:

1. *agonistic fairness*: promotes exploration
2. *antagonistic fairness*: inhibits exploration.

Fairness Constraints:

(i) *Pre-specified constraints*: we select the fairness vector as

$$p_a = p_0[\alpha w_a^\star + (1-\alpha)\bar{w}_a^\star], \quad \bar{w}_a^\star = (1/w_a^\star)/\sum_{b\in[K]}(1/w_b^\star).$$

where $w^\star$ is optimal for $T_0^\star$. We set $\alpha = 0.9$ for the *agonistic* case and $\alpha = 0.1$ in the *antagonistic* one.

(ii) *θ-dependent constraints:*

- In the *agonistic case* we select the fairness functions as $p_a(\theta) = p_0 \frac{1/\max(\Delta_a,\Delta_{\min})}{\sum_{b\in[K]} 1/\max(\Delta_b,\Delta_{\min})}$
- In the *antagonistic case* we select $p_a(\theta) = p_0 \frac{\Delta_a}{\sum_{b\in[K]}\Delta_b}$.

Wireless Scheduling: Fairness

Settings: we focus on two settings to analyze how p impacts exploration:

1. *agonistic fairness*: promotes exploration
2. *antagonistic fairness*: inhibits exploration.

Fairness Constraints:

(i) *Pre-specified constraints*: we select the fairness vector as

$$p_a = p_0[\alpha w_a^\star + (1-\alpha)\bar{w}_a^\star], \quad \bar{w}_a^\star = (1/w_a^\star)/\sum_{b\in[K]}(1/w_b^\star).$$

where $w^\star$ is optimal for $T_0^\star$. We set $\alpha = 0.9$ for the *agonistic* case and $\alpha = 0.1$ in the *antagonistic* one.

(ii) *θ-dependent constraints:*

- In the *agonistic case* we select the fairness functions as $p_a(\theta) = p_0 \frac{1/\max(\Delta_a, \Delta_{\min})}{\sum_{b\in[K]} 1/\max(\Delta_b, \Delta_{\min})}$
- In the *antagonistic case* we select $p_a(\theta) = p_0 \frac{\Delta_a}{\sum_{b\in[K]}\Delta_b}$.

Wireless Scheduling: Fairness

Settings: we focus on two settings to analyze how p impacts exploration:

1. *agonistic fairness*: promotes exploration
2. *antagonistic fairness*: inhibits exploration.

Fairness Constraints:

(i) *Pre-specified constraints*: we select the fairness vector as

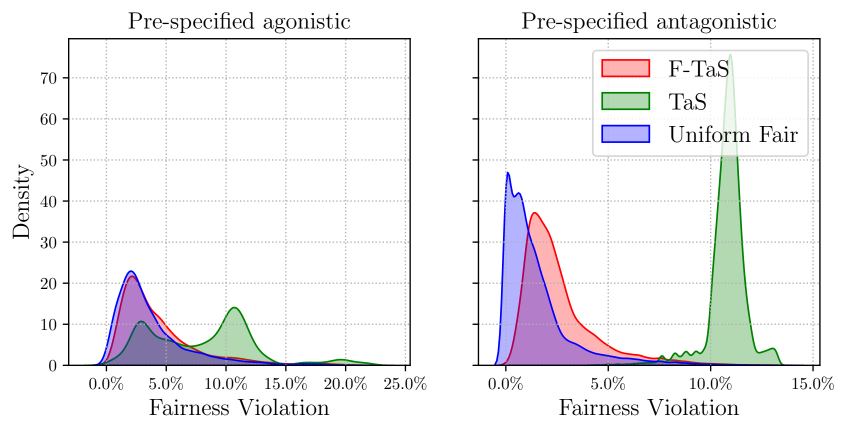
$$p_a = p_0[\alpha w_a^\star + (1-\alpha)\bar{w}_a^\star], \quad \bar{w}_a^\star = (1/w_a^\star)/\sum_{b\in[K]}(1/w_b^\star).$$

where $w^\star$ is optimal for $T_0^\star$. We set $\alpha = 0.9$ for the *agonistic* case and $\alpha = 0.1$ in the *antagonistic* one.

(ii) *θ-dependent constraints:*

- In the *agonistic case* we select the fairness functions as $p_a(\theta) = p_0 \frac{1/\max(\Delta_a, \Delta_{\min})}{\sum_{b\in[K]} 1/\max(\Delta_b, \Delta_{\min})}$
- In the *antagonistic case* we select $p_a(\theta) = p_0 \frac{\Delta_a}{\sum_{b\in[K]}\Delta_b}$.

Wireless Scheduling: Fairness Violation, Pre-specified scenario

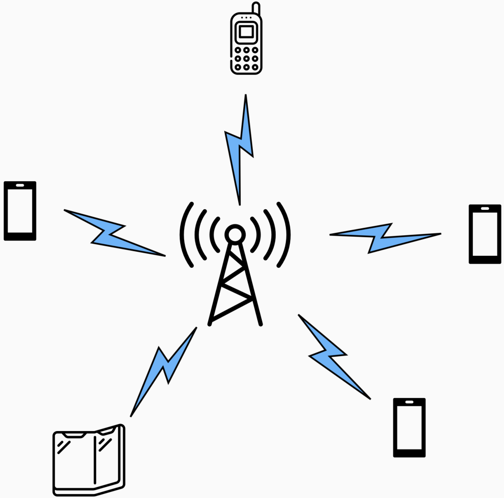


Fairness violation: $\rho(t) = \max(0, \max_a p_a(\theta) - N_a(t)/t)$.

Wireless Scheduling: sample complexity results

	Algorithm	Pre-specified constraints				θ-dependent constraints			
		Sample Complexity		Fairness Violation		Sample Complexity		Fairness Violation	
		Agonistic	Antagonistic	Agonistic	Antagonistic	Agonistic	Antagonistic	Agonistic	Antagonistic
$\delta = 0.1$	F-TaS	199.10 ± 15.96	457.90 ± 48.15	$3.03\% \pm 0.39\%$	$2.13\% \pm 0.24\%$	197.80 ± 17.05	599.79 ± 68.83	$4.60\% \pm 0.43\%$	$2.97\% \pm 0.32\%$
	TaS	136.88 ± 9.59	136.88 ± 9.78	$6.55\% \pm 0.68\%$	$10.76\% \pm 0.12\%$	136.88 ± 9.48	136.88 ± 9.86	$5.32\% \pm 0.36\%$	$8.22\% \pm 0.08\%$
	Uniform Fair	236.50 ± 16.11	726.52 ± 85.13	$2.45\% \pm 0.37\%$	$1.12\% \pm 0.25\%$	220.07 ± 18.00	1889.56 ± 287.37	$4.07\% \pm 0.35\%$	$1.94\% \pm 0.48\%$
$\delta = 0.01$	F-TaS	285.41 ± 15.74	696.11 ± 58.62	$2.35\% \pm 0.27\%$	$1.79\% \pm 0.20\%$	298.68 ± 21.88	833.55 ± 78.24	$3.96\% \pm 0.37\%$	$2.38\% \pm 0.23\%$
	TaS	207.79 ± 13.53	207.79 ± 13.64	$5.71\% \pm 0.67\%$	$11.14\% \pm 0.13\%$	207.79 ± 13.84	207.79 ± 13.28	$4.92\% \pm 0.37\%$	$8.55\% \pm 0.11\%$
	Uniform Fair	323.86 ± 19.23	1071.62 ± 91.97	$1.91\% \pm 0.29\%$	$0.68\% \pm 0.18\%$	359.49 ± 24.66	2853.99 ± 319.41	$3.00\% \pm 0.26\%$	$1.21\% \pm 0.40\%$
$\delta = 0.001$	F-TaS	358.81 ± 17.44	899.13 ± 74.28	$2.00\% \pm 0.29\%$	$1.60\% \pm 0.18\%$	398.94 ± 24.53	1048.52 ± 84.89	$3.43\% \pm 0.34\%$	$2.02\% \pm 0.18\%$
	TaS	271.05 ± 16.99	271.05 ± 16.87	$5.22\% \pm 0.62\%$	$11.51\% \pm 0.10\%$	271.05 ± 16.93	271.05 ± 17.11	$4.67\% \pm 0.33\%$	$8.90\% \pm 0.10\%$
	Uniform Fair	410.72 ± 22.63	1383.06 ± 95.08	$1.52\% \pm 0.21\%$	$0.41\% \pm 0.12\%$	476.13 ± 32.11	3703.97 ± 354.92	$2.58\% \pm 0.24\%$	$0.86\% \pm 0.37\%$

Motivation



Caption